Express Mail EG365635465US
Mailing Date: June 18, 1998

Patent Application
Attorney Docket No.: D95465

WHAT IS CLAIMED IS:

a
a

1. A processor implemented method of identifying a ~~text genre~~ document type of ~~an~~ a document ~~untagged text~~ document in machine-readable form without structurally analyzing the document text, the processor implemented method comprising the steps of:

a) generating a cue vector from the text, the cue vector representing occurrences in the text of a first set of nonstructural, surface cues; and

b) determining whether the text is an instance of a first text genre using the cue vector and a weighting vector associated with the first text genre.

2. The method of claim 1 wherein the first set of cues includes a punctuational cue.

3. The method of claim 2 wherein the punctuational cue represents a one of a number of commas in the text, a number of dashes in the text, a number of question marks in the text and a number of semi-colons in the text.

4. The method of claim 1 wherein the first set of cues includes a string recognizable constructional cue.

a

5. The method of claim ~~1~~ 4 wherein the string recognizable constructional cue represents a one of a first number of sentences starting with the words “and”, “but” and “so” and a second number of sentences starting with an adverb and a comma.

6. The method of claim 1 wherein the first set of cues includes a formulae cue.

7. The method of claim 1 wherein the first set of cues includes a lexical cue.

8. The method of claim 7 wherein the lexical cue represents a one of a first number of occurrences in the text of acronyms, a second number of occurrences in the text of modal auxiliaries, a third number of occurrences of form of the verb “be”, and a fourth number of occurrences of calendar words.

9. The method of claim 7 wherein the lexical cue represents a one of a first number of occurrences in the text of capitalized words, a second number of occurrences in the text of contractions, a third number of occurrences in the text of words that end in “ed”, and a fourth number of occurrences in the text of mathematical formulas.

10. The method of claim 7 wherein the lexical cue represents a one of a first number of occurrences in the text of polysyllabic words, a second number of occurrences in the text of the word “it”, a third number of occurrences in the text of latinate prefixes and suffixes, and a fourth number of occurrences in the text of overt negatives.

11. The method of claim 7 wherein the lexcial cue represents a one of a first number of occurrences in the text of words including at least one digit, a second number of occurrences in the text of left parentheses, a third number of occurrences in the text of prepositions, a fourth number of occurrences in the text of first person pronouns, and a fifth number of occurrences in the text of second person pronouns.

12. The method of claim 7 wherein the lexical cue represents a one of a first number of occurrences in the text of quotation marks, a second number of occurrences in the text of roman numerals, a third number of occurrences in the text of “that”, and a fourth number of occurrences in the text of “which”.

13. The method of claim 2 wherein the first set of cues includes a deviation cue.

14. The method of claim 13 wherein the deviation cue includes a one of a first deviation of a sentence length of the text and a second deviation of a word length of the text.

15. The method of claim 3 wherein the first set of cues further includes a second set of lexical cues, a third set of string recognizable constructional cues, a fourth set of formulae cues and fifth set of deviation cues.

16. The method of claim 15 wherein the second set of lexical cues includes at least a one lexical cue representing a one of a first number of occurrences in the text of acronyms, a second number of occurrences in the text of modal auxiliaries, a third number of occurrences of form of the verb “be”, a fourth number of occurrences of calendar words, a fifth number of occurrences in the text of capitalized words, a sixth number of occurrences in the text of contractions, a seventh number of occurrences in the text of words that end in “ed, an eighth number of occurrences in the text of mathematical formulas, a ninth number of occurrences in the text of polysyllabic words, a tenth number of occurrences in the text of the word “it”, an eleventh number of occurrences in the text of Latinate prefixes and suffixes, a twelfth number of occurrences in the text of overt negatives, a thirteenth number of occurrences in the text of words including at least one digit, a fourteenth number of occurrences in the text of parentheses, a fifteenth number of occurrences in the text of prepositions, a sixteenth number of occurrences in the text of first person pronouns, a seventeenth number of occurrences in the text of second person pronouns, an eighteenth number of occurrences in the text of quotation marks, a nineteenth number of occurrences in the text of roman numerals, a twentieth number of

occurrences in the text of "that", and a twenty-first number of occurrences in the text of "which".

17. The method of claim 15 wherein the third set of string recognizable constructional cues includes at least one string recognizable constructional cue representing a one of a first number of sentences starting with the words "and", "but" and "so" and a second number of sentences starting with an adverb and a comma.

18. The method of claim 15 wherein the fifth set of deviation cues includes at least one deviation cue representing a one of a first deviation of a sentence length of the text and a second deviation of a word length of the text.

Sub a1

~~19. A processor implemented method of identifying a text genre of an untagged text in machine-readable form without structurally analyzing the text, the processor implemented method comprising the steps of:~~

~~a) generating a cue vector from the text, the cue vector representing occurrences in the text of a first set of nonstructural, surface cues;~~

~~b) determining a relevancy to the text of each facet of a second set of facets using the cue vector and a weighting vector; and~~

~~c) identifying from a third set of text genre types a text genre type of the text based upon those facets of the second set that are relevant to the text.~~

20. The method of claim 19 wherein the first set of cues includes a punctuational cue.

21. The method of claims 19 wherein the first set of cues includes a one of includes a lexical cue, a string recognizable constructional cue, a formulae cue and a deviation cue.

22. The method of claim 19 wherein the second set of facets includes at least a one of a date facet, a narrative facet, a suasive facet, a fiction facet, a legal facet ~~fact~~, a science and technical facet, and an author facet.

a

23. The method of claim 19 wherein the third set of text genre types includes at least a one of a press report type, an Email type, an editorial opinion type, and a market analysis type.

24. The method of claim 21 wherein the second set of facets includes at least a one of a date facet, a narrative facet, a suasive facet, a fiction facet, a legal facet, a science and technical facet, and an author facet.

25. The method of claim 24 wherein the third set of text genre types includes at least a one of a press report type, an Email type, an editorial opinion type, and a market analysis type.

Sub a7

~~26. An article of manufacture comprising:~~

~~a) a memory; and~~

~~b) instructions stored in the memory for a method of identifying a text genre of an untagged text in machine-readable form without structurally analyzing the text, the method being implemented by a processor coupled to the memory, the instructions comprising the steps of:~~

~~1) generating a cue vector from the text, the cue vector representing occurrences in the text of a first set of nonstructural, surface cues; and~~

~~2) determining whether the text is an instance of a first text genre using the cue vector and a weighting vector associated with the first text genre.~~

~~27. An article of manufacture comprising:~~

~~a) a memory; and~~

~~b) instructions stored in the memory for a method of identifying a text genre of an untagged text in machine-readable form without structurally analyzing the text, the method being implemented by a processor coupled to memory, the instructions comprising the steps of:~~

~~1) generating a cue vector from the text, the cue vector representing occurrences in the text of a first set of nonstructural, surface cues, the first set of cues including a punctuational cue;~~

~~2) determining a relevancy to the text of each facet of a second set of facets using the cue vector and a weighting vector; and~~

~~3) identifying from a third set of text genre types a text genre type of the text based upon those facets of the second set that are relevant to the text.~~